



UNITED STATES PATENT AND TRADEMARK OFFICE

UNITED STATES DEPARTMENT OF COMMERCE
United States Patent and Trademark Office
Address: COMMISSIONER FOR PATENTS
P.O. Box 1450
Alexandria, Virginia 22313-1450
www.uspto.gov

APPLICATION NO.	FILING DATE	FIRST NAMED INVENTOR	ATTORNEY DOCKET NO.	CONFIRMATION NO.
10/634,616	08/04/2003	Alexander Franz	025.0289.US.UTL	6590

22895 7590 02/05/2007
CASCADIA INTELLECTUAL PROPERTY
500 UNION STREET
SUITE 1005
SEATTLE, WA 98101

EXAMINER

HERNANDEZ, JOSIAH J

ART UNIT	PAPER NUMBER
----------	--------------

2609

SHORTENED STATUTORY PERIOD OF RESPONSE	MAIL DATE	DELIVERY MODE
3 MONTHS	02/05/2007	PAPER

Please find below and/or attached an Office communication concerning this application or proceeding.

If NO period for reply is specified above, the maximum statutory period will apply and will expire 6 MONTHS from the mailing date of this communication.

Office Action Summary

Application No.

10/634,616

Applicant(s)

FRANZ ET AL.

Examiner

Josiah Hernandez

Art Unit

2609

-- The MAILING DATE of this communication appears on the cover sheet with the correspondence address --

Period for Reply

A SHORTENED STATUTORY PERIOD FOR REPLY IS SET TO EXPIRE 3 MONTH(S) OR THIRTY (30) DAYS, WHICHEVER IS LONGER, FROM THE MAILING DATE OF THIS COMMUNICATION.

- Extensions of time may be available under the provisions of 37 CFR 1.136(a). In no event, however, may a reply be timely filed after SIX (6) MONTHS from the mailing date of this communication.
- If NO period for reply is specified above, the maximum statutory period will apply and will expire SIX (6) MONTHS from the mailing date of this communication.
- Failure to reply within the set or extended period for reply will, by statute, cause the application to become ABANDONED (35 U.S.C. § 133). Any reply received by the Office later than three months after the mailing date of this communication, even if timely filed, may reduce any earned patent term adjustment. See 37 CFR 1.704(b).

Status

- 1) ☐ Responsive to communication(s) filed on 04 August 2003.
- 2a) ☐ This action is **FINAL**. 2b) ☒ This action is non-final.
- 3) ☐ Since this application is in condition for allowance except for formal matters, prosecution as to the merits is closed in accordance with the practice under *Ex parte Quayle*, 1935 C.D. 11, 453 O.G. 213.

Disposition of Claims

- 4) ☒ Claim(s) 37 is/are pending in the application.
- 4a) Of the above claim(s) _____ is/are withdrawn from consideration.
- 5) ☐ Claim(s) _____ is/are allowed.
- 6) ☒ Claim(s) 1-37 is/are rejected.
- 7) ☐ Claim(s) _____ is/are objected to.
- 8) ☐ Claim(s) _____ are subject to restriction and/or election requirement.

Application Papers

- 9) ☐ The specification is objected to by the Examiner.
- 10) ☒ The drawing(s) filed on 04 August 2003 is/are: a) ☒ accepted or b) ☐ objected to by the Examiner.
- Applicant may not request that any objection to the drawing(s) be held in abeyance. See 37 CFR 1.85(a).
- Replacement drawing sheet(s) including the correction is required if the drawing(s) is objected to. See 37 CFR 1.121(d).
- 11) ☐ The oath or declaration is objected to by the Examiner. Note the attached Office Action or form PTO-152.

Priority under 35 U.S.C. § 119

- 12) ☐ Acknowledgment is made of a claim for foreign priority under 35 U.S.C. § 119(a)-(d) or (f).
- a) ☐ All b) ☐ Some * c) ☐ None of:
- ☐ Certified copies of the priority documents have been received.
 - ☐ Certified copies of the priority documents have been received in Application No. _____.
 - ☐ Copies of the certified copies of the priority documents have been received in this National Stage application from the International Bureau (PCT Rule 17.2(a)).

* See the attached detailed Office action for a list of the certified copies not received.

Attachment(s)

- 1) ☒ Notice of References Cited (PTO-892)
- 2) ☐ Notice of Draftsperson's Patent Drawing Review (PTO-948)
- 3) ☒ Information Disclosure Statement(s) (PTO/SB/08)
Paper No(s)/Mail Date 10/27/2003.
- 4) ☐ Interview Summary (PTO-413)
Paper No(s)/Mail Date. _____.
- 5) ☐ Notice of Informal Patent Application
- 6) ☐ Other: _____.

DETAILED ACTION

Claim Rejections - 35 USC § 103

1. The following is a quotation of 35 U.S.C. 103(a) which forms the basis for all obviousness rejections set forth in this Office action:

(a) A patent may not be obtained though the invention is not identically disclosed or described as set forth in section 102 of this title, if the differences between the subject matter sought to be patented and the prior art are such that the subject matter as a whole would have been obvious at the time the invention was made to a person having ordinary skill in the art to which said subject matter pertains. Patentability shall not be negated by the manner in which the invention was made.

2. Claims 1,2,4,6,9-16,18,20,23-37 are rejected under 35 U.S.C. 103(a) as being unpatentable over van den Akker (US 6,415,250) in view of Bracewell et al. (US PGPub 2006/0041685).

As to claims 1 and 15, van den Akker discloses a system for identifying language attributes through probabilistic analysis (see abstract lines 1-3; column 3 paragraph 1; column 7 paragraph 4). Van den Akker also discloses a storage storing a set of language classes, which identify a language and a character set encoding (see column 3 paragraph 2; column 9 paragraph 2; column 11 lines 34-39), and a plurality of training documents (see column 9 lines 5-8, 42-46; column 10 lines 5-10). Van den Akker discloses a text modeler evaluating byte

occurrence within each training document and, for each language class, calculating a probability for the byte occurrences conditioned on the occurrence of the language class (see column 3 lines 15-20, 25-35; column 9 lines 15-20). Van den Akker does not specifically disclose having an attribute modeler that evaluates occurrences of one or more document properties within each training document and, for each language class, calculating a probability for the document properties set conditioned on the occurrence of the language class. Bracewell teaches that document properties like HTTP header information can be used in order to identify the language of a document or search query on the internet (see [0014]). It would have been obvious to have used the attribute properties, such as the HTTP header information, as disclosed in Bracewell in a language identification model taught by van den Akker because using text and document property will result in an efficient form of identifying the language of a document or search query (see [0014]).

As to claims 2 and 16, van den Akker discloses a training engine calculating an overall probability for each language class by evaluating the probability for the document property set and the probability for the byte occurrences (see column 3 lines 35-45, 64-67; column 4 lines 50-55; column 10 lines 33-40).

As to claims 4 and 18, van den Akker does not disclose specifically that the document properties comprise at least one of top level domain, HTTP content character set encoding and language header parameters, and HTML content character set encoding and language metatags. Bracewell teaches that document properties like HTTP header information can be used in order to identify the language of a document or search query on the Internet (see [0014]). It would have been obvious to have used the attribute properties, such as the HTTP header information, as disclosed in Bracewell in a language identification model taught by van den Akker because using text and document property will result in an efficient form of identifying the language of a document or search query (see [0014]).

As to claims 6 and 20, van den Akker discloses a counting module counting byte co-occurrences within each training document, and determining the probability for the byte occurrences based on the byte co-occurrences (see column 3 lines 15-20, 50-55; column 12 lines 50-55).

As to claims 9 and 23, van den Akker discloses a training engine performing iterative training by providing the probability for the document properties set and the probability for the byte occurrences set respectively to the

evaluation of byte occurrences and assignment of the set of language classes
(see column 3 lines 35-45, 64-67; column 4 lines 50-55; column 10 lines 33-40).

As to claims 10 and 24, van den Akker discloses a back off module evaluating less frequently occurring document properties by calculating a probability for each less frequently occurring document property conditioned on the occurrence of the language class (see column 10 45-65).

As to claims 11 and 25, van den Akker discloses a plurality of unlabeled documents (see column 7 lines 50-55) and a classifier classifying one or more unlabeled documents by at least one language class (see column 5 lines 36-44), comprising evaluating occurrences of one or more document properties within the unlabeled document, evaluating byte occurrences within the unlabeled document, and assigning a probability for the document properties set and the byte occurrences for the unlabeled document conditioned on the occurrence of the language class (see column 7 lines 55-67)

As to claims 12 and 26, van den Akker discloses a class selector selecting at least one language class having a substantially highest probability (see column 5 paragraph 2 lines 23-26).

As to claims 13 and 24, van den Akker discloses a probability threshold and a pruner pruning at least one language class falling below the probability threshold (see column 13 paragraph 2 lines 17-25).

As to claims 14 and 28, van den Akker discloses training document comprising one of a Web page and a news message (see column 4 lines 15-17; column 5 lines 27-37; column 7 lines 56-62).

As to claims 29 and 36, van den Akker discloses a computer-readable storage medium holding code for performing the method of identifying language attributes through probabilistic analysis (see column 6 lines 39-60).

As to claims 30 and 33, van den Akker discloses a system for identifying documents by language using probabilistic analysis of language attributes (see abstract lines 1-3; column 3 paragraph 1; column 7 paragraph 4), comprising: a set of language classes, each language class comprising a language and a character set encoding name (see column 3 paragraph 2; column 9 paragraph 2; column 11 lines 34-39); a training corpora comprising a plurality of training documents (see column 9 lines 5-8, 42-46; column 10 lines 5-10); and a text modeler training a text model by evaluating co-occurrences of a plurality of bytes within each training document and , for each language class, calculating a probability for the byte co-occurrences conditioned on the occurrence of the each

language class (see column 3 lines 15-20, 25-35; column 9 lines 15-20). Van den Akker does not specifically disclose an attribute modeler training an attribute model by evaluating a top level domain and character set encoding associated with each training document and, for each language class, calculating a probability for each such top level domain and character set encoding conditioned on the occurrence of the each language class. Bracewell teaches that document properties like HTTP header information can be used in order to identify the language of a document or search query on the internet (see [0014]). It would have been obvious to have used the attribute properties, such as the HTTP header information, as disclosed in Bracewell in a language identification model taught by van den Akker because using text and document property will result in an efficient form of identifying the language of a document or search query (see [0014]).

As to claims 31 and 34, van den Akker discloses a training engine calculating an overall probability for each language class by evaluating the probability for the top level domain and character set encoding based on the attribute model and the probability for the byte occurrences based on the text model (see column 3 lines 35-45, 64-67; column 4 lines 50-55; column 10 lines 33-40).

As to claim 32, van den Akker discloses a classifier classifying one or more documents (see column 5 lines 36-44), comprising: a text evaluator evaluating byte occurrences in each document and applying the text model to the evaluated byte occurrences (see column 3 lines 15-20, 25-35; column 9 lines 15-20); and an assignment module assigning at least one language class based on the applications of the attribute model and the text model (see column 5 lines 24-26). Van den Akker does not specifically disclose an attribute evaluator evaluating a top level domain and character set encoding in each document and applying the attribute model to the evaluated top level domain and character set encoding. Bracewell teaches that document properties like HTTP header information can be used in order to identify the language of a document or search query on the internet (see [0014]). It would have been obvious to have used the attribute properties, such as the HTTP header information, as disclosed in Bracewell in a language identification model taught by van den Akker because using text and document property will result in an efficient form of identifying the language of a document or search query (see [0014]).

As to claim 35, van den Akker discloses classifying one or more documents (see column 5 lines 36-44), comprising: evaluating byte occurrences in each document and applying the text model to the evaluated byte occurrences (see column 3 lines 15-20, 25-35; column 9 lines 15-20); and assigning at least

one language class based on the applications of the attribute model and the text model (see column 5 lines 24-26). Van den Akker does not specifically disclose evaluating a top level domain and character set encoding in each document and applying the attribute model to the evaluated top level domain and character set encoding. Bracewell teaches that document properties like HTTP header information can be used in order to identify the language of a document or search query on the internet (see [0014]). It would have been obvious to have used the attribute properties, such as the HTTP header information, as disclosed in Bracewell in a language identification model taught by van den Akker because using text and document property will result in an efficient form of identifying the language of a document or search query (see [0014]).

As to claim 37, van den Akker discloses an apparatus for identifying documents by language using probabilistic analysis of language attributes (see abstract lines 1-3; column 3 paragraph 1; column 7 paragraph 4), comprising: means for defining a set of language classes, each language class comprising a language name and a character set encoding name (see column 3 paragraph 2; column 9 paragraph 2; column 11 lines 34-39); means for training a text model by evaluating co-occurrences of a plurality of bytes within each training document, for each language class, calculating a probability for the byte co-occurrence conditioned on the occurrence of the language class based on the

attribute model (see column 3 lines 15-20, 25-35; column 9 lines 15-20). Van den Akker does not disclose specifically means for training an attribute model by assigning at least one top level domain and character set encoding pairing to at least one language class for each of a plurality of training documents and calculating a probability for each such top level domain and character set encoding pairing conditioned on the occurrence of the assigned language class. Bracewell teaches that document properties like HTTP header information can be used in order to identify the language of a document or search query on the internet (see [0014]). It would have been obvious to have used the attribute properties, such as the HTTP header information, as disclosed in Bracewell in a language identification model taught by van den Akker because using text and document property will result in an efficient form of identifying the language of a document or search query (see [0014]).

3. Claims 3, 5, 17, and 19 are rejected under 35 U.S.C. 103(a) as being unpatentable over van den Akker (US 6,415,250) in view of Bracewell et al. (US PGPub 2006/0041685) as applied to claims 1,2,4,6,9-16,18,20,23-37 above, and in further view of Elworthy (US 6,125,362).

As to claims 3 and 17, van den Akker and Bracewell do not specifically disclose an assignment module assigning the overall probability for each language class in accordance with the formula: $\arg \max P(\text{text}|\text{cls}) * P(\text{props}|\text{cls}) * P(\text{cls})$. Elworthy teaches the use of probabilistic analysis for determining the language of a text or document (see column 1 lines 37-40). Elworthy further teaches that in order to classify documents or text according to a certain element the following Bayesian probabilistic formula can be used: $p(l|t) = (P(t|l) * p(l)) / p(t)$ (see column 4 lines 25-35). If the denominator is passed to the left side the resultant equation is: $p(l|t) * p(t) = P(t|l) * p(l)$, where $p(l|t)$ is the probability of the classification given the element and $p(t)$ is the probability of the element (see column 4 lines 35-45). It would have been obvious to have used $p(l|t) * p(t)$ as disclosed in Elworthy for the probabilistic analysis in van den Akker as modified, where $p(l|t) = P(\text{text}|\text{cls}) * P(\text{props}|\text{cls})$, where t is the language class and l is the text or the attribute property and $p(l|t) * p(t)$ would be the probability of the language class given the text and the attribute. It would have been obvious to have used both prior arts because using the method described above would yield an accurate form of identifying the language model (see column 4 lines 25-45).

As to claims 5 and 19, van den Akker and Bracewell do not specifically disclose an assignment module assigning the probability for the document properties set in accordance with the formula: $P(\text{tld,enc}|\text{cls}) * P(\text{cls})$. Elworthy

teaches the use of probabilistic analysis for determining the language of a text or document (see column 1 lines 37-40). Elworthy further teaches that in order to classify documents or text according to a certain element the following Bayesian probabilistic formula can be used: $p(I|t) = (P(t|I)*p(I))/p(t)$ (see column 4 lines 25-35). If the denominator is passed to the left side the resultant equation is: $p(I|t) * p(t) = P(t|I)*p(I)$, where $p(I|t)$ is the probability of the classification given the element and $p(t)$ is the probability of the element (see column 4 lines 35-45). It would have been obvious to have used $p(I|t)*p(t)$ as disclosed in Elworthy for the probabilistic analysis in van den Akker as modified, where $p(I|t) = P(t|I,enc|cls)*P(cls)$, where t is the language class and I is the text or the attribute property and $p(I|t)*p(t)$ would be the probability of the language class given the text and the attribute. It would have been obvious to have used both prior arts because using the method described above would yield an accurate form of identifying the language model (see column 4 lines 25-45).

4. Claims 7 and 21 are rejected under 35 U.S.C. 103(a) as being unpatentable over van den Akker (US 6,415,250) in view of Bracewell et al. (US PGPub 2006/0041685) as applied to claims 1,2,4,6,9-16,18,20,23-37 above, and in further view of de Campos (US 6,272,456).

As to claims 7 and 21, van den Akker and Bracewell do not specifically disclose using trigrams. De Campos teaches using the byte co-occurrences comprise a set of trigrams (see column 1 lines 59-67; column 2 lines 50-54, 59-64; column 6 lines 53-60), further comprising a probability module calculating a probability of each trigram as the number of occurrences of the trigram divided by the total number of trigram occurrences in each of the training documents for each language class (see column 18 lines 64-67 and column 19 lines 1-6). It would have been obvious to use the trigram method disclosed in de Campos for the byte co-occurrences in the text model in van den Akker as modified because the trigram method would allow the text model to break down the unlabeled text and identify the language (see column 18 lines 64-67 and column 19 lines 1-6).

5. Claims 8 and 22 are rejected under 35 U.S.C. 103(a) as being unpatentable over van den Akker (US 6,415,250) in view of Bracewell et al. (US PGPub 2006/0041685) as applied to claims 1,2,4,6,9-16,18,20,23-37 above, and in further view of de Campos (US 6,272,456) and Elworthy (US 6,125,362).

As to claims 8 and 22, van den Akker, Bracewell and Campos do not specifically disclose an assignment module assigning the overall probability for each language class in accordance with the formula: $P(\text{text}|\text{cls})$. Elworthy teaches the use of probabilistic analysis for determining the language of a text or

document (see column 1 lines 37-40). Elworthy further teaches that in order to classify documents or text according to a certain element the following Bayesian probabilistic formula can be used: $p(l|t) = (P(t|l)*p(l))/p(t)$ (see column 4 lines 25-35). If the denominator is passed to the left side the resultant equation is: $p(l|t) * p(t) = P(t|l)*p(l)$, where $p(l|t)$ is the probability of the classification given the element and $p(t)$ is the probability of the element (see column 4 lines 35-45). It would have been obvious to have used $p(l|t)$ as disclosed in Elworthy for the probabilistic analysis in van den Akker as modified, where $p(l|t) = P(\text{text}|cls)$, where t is the language class and l is the text or set of trigrams (as disclosed by de Campos) and $p(l|t)$ would be the probability of the language class given the text or trigram. It would have been obvious to have used both prior arts because using the method described above would yield an accurate form of identifying the language model (see column 4 lines 25-45).

Conclusion

Any inquiry concerning this communication should be directed to Josiah Hernandez whose telephone number is 571-270-1646. The examiner can normally be reached from 7:30 pm to 5:00 pm.

If attempts to reach the examiner by telephone are unsuccessful, the examiner's supervisor, Xiao Wu can be reached on (571) 272-7761. The fax

Art Unit: 2609

phone number for the organization where this application or proceeding is assigned is 703-872-9306.

Information regarding the status of an application may be obtained from the Patent Application Information Retrieval (PAIR) system. Status information for published applications may be obtained from either Private PAIR or Public PAIR. Status information for unpublished applications is available through Private PAIR only. For more information about the PAIR system, see <http://pair-direct.uspto.gov>. Should you have questions on access to the Private PAIR system, contact the Electronic Business Center (EBC) at 866-217-9197 (toll-free).

J.H.

1/25/2007


XIAO WU
SUPERVISORY PATENT EXAMINER